

**NWX-US DEPT OF COMMERCE**

**Moderator: Deborah Rivera Nieves**  
**June 3, 2019**  
**1:00 pm CT**

Coordinator: Welcome and thank you for standing by. At this time all participants are in a listen-only mode until the question-and-answer session of today's conference. At that time, you may press Star followed by 1 on your phone to ask a question. I would like to inform all parties that today's conference is being recorded. If you have any objections you may disconnect at this time. I would now like to turn the conference over to Deborah Rivera. Thank you. You may begin.

Deborah Rivera: Great, thank you so much (Marcus) for that introduction. Good afternoon, everyone. As (Marcus) stated, I am Deborah Rivera, your host for today. And I would first like to welcome all of you who joined our Census Academy webinar, our overview of the 2014 panel of the Survey of Income and Program Participation. This is the first of seven webinars that will be taking place during the month of June. And it aims to give our data users a thorough look into the Survey of Income and Program Participation from the survey design itself, to the many topics that it collects data on.

Today's webinar will be an overview as the title suggests, and it covers the survey content, the design and processing. And you are welcome to join us

tomorrow for Part 2 of the SIPP webinar series where our presenter Holly Fee will be covering demographics and residences. And also as an FYI, we do have course material available for this webinar and tomorrow's webinar on the Census Academy site.

So a few housekeeping items before we get started here. We are recording this webinar. And along with the PowerPoint slides, we will be posting that recording to the Census Academy site as a free learning resource. We will hold off for questions until the end of the presentation. At which time we ask that you keep to one question. And if you have any questions remaining after that, you're welcome to queue back up to ask again.

I would now like to introduce our speaker for today, Mr. Matthew Marlay. Matthew Marlay is the Chief of the SIPP Coordination and Outreach Staff and he has been at the Census Bureau since 2008 and has worked on SIPP that entire time in a variety of capacities. Prior to joining the Census Bureau, Matthew received a Ph.D. in sociology and demography from Penn State. Thank you so much, Matthew.

Matthew Marlay: Thank you, Deborah. Welcome everybody, we're really happy that you're. We're really pleased to be doing this webinar series. We had, sort of, been planning one and then Deborah's team reached out to us and asked us if we could done one. And we said well why don't we just do a whole series of them? So they got seven despite only asking for one.

So as Deborah said, today' is the overview. This will cover sort of a high level, all different facets of the 2014 SIPP. If you're interested in more specific subject matter, we have a whole series of them coming up throughout the month of June as you can see here. And tomorrow's will be demographics and residences that Holly will present. So we do have Holly Fee and Shelley

Irving on the line. They are also in the SIPP Coordination and Outreach Staff and they'll be answering questions and handling the chat and things like that.

So again, before we get started, the slides are available as Deb said on the Census Academy site here. And there are also some supplemental materials that are available as well. There's some sample code. The more topic heavy webinars will have a lot more supplemental materials. For the opening one, it's mostly just the slides, but there is a little bit of sample code as well.

So here's what we're going to cover today. I'll start by talking about the background and the history of the SIPP. I'll give a little overview of the survey and sample design as well as talking about the content and how we process the data. And then just some things to be aware of when you're using the data. And then we'll finish with some resources for you guys to look at if you have questions or need to get ahold of us after the webinar.

So just to start with a little bit of background about SIPP. SIPP's mission is to provide a nationally representative sample that allows researchers to evaluate things such as the annual and semi-annual dynamics of income. And for SIPP that really means monthly income. We have income to that granular of a level. You can also look at movement in and out of government transfer programs. You can look at the family and social context of individuals and households as well as interactions between these items.

So SIPP is a really rich dataset and as such, it has a number of advantages. It's got comprehensive detailed data on a whole lot of different topics, and you'll get a preview of that throughout this webinar. You can also look at the dynamics of different levels. You can look at household. You can look at families. You can look at individuals depending on your research interests. SIPP is longitudinal over about a four year period. The 2014 panel ran through

2017, and SIPP follows movers. And so, again, I'll talk a little bit more about that. But it allows you to, sort of, see how people's circumstances change over a four-year period.

So SIPP began in 1984 and we ran panels through 1993 with an overlapping sample - overlapping panel design. So in other words, the 1984 panel hadn't finished when the 1985 panel started and so on. Starting in 1996, we converted it from a paper instrument to a DOS-based instrument. And at that point the panels were no longer overlapping. They were end-to-end. So 1996 panel ran until 2000. Then a new panel started in 2001, and so on.

With the 2014 panel, we debuted a totally reworked instrument with different - mostly the same content, but some content differences. Some significant changes to the instruments and the data processing. And a lot of that is what I'll talk about today. That panel ran through 2017 and has now been concluded. So if you've used SIPP before, if you're familiar with how SIPP used to be designed, you know that there was a four-month recall period meaning that we did three interviews of each household per year.

We are always in the field. So the sample was divided into four rotation groups. And each of those rotation groups was interviewed monthly. There were some questions that were asked every interview. And we refer to that as core content. And then other questions that were asked less frequently on very specific subject matter areas and those or what's referred to as topical module content.

So with the current SIPP, we've redesigned it. So now it's an annual interview. So we're asking people about the previous twelve-months instead of the previous three months, four months. And we have introduced an event history calendar to, sort of, help with respondent's recall. And here's what that looks

like. We'll be talking about the event history calendar a lot. Will also refer to it as the EHC. And so here is just to give you an idea of what that looks like. And you can see that there are months along the top, and there are topics in each row. And really what this allows you to do is to go through the interview in any order that makes sense for the respondent.

So if for example, we're talking about the resident's history and we're going to ask about this whole year's worth of resident's history, and the respondent says I moved in June. Oh, that's because I got a new job. The field representative or our interviewer, those are the same thing, can immediately then go down to the job line because the respondent has started talking about their job and collect the job information right then. And similarly, if the respondent says oh I changed jobs and that caused my health insurance to change, again, they can come down here to the health insurance line and start entering that information right away.

Now in practice, people do tend to go through the event history calendar from top to bottom. And, you know, they start - we usually start in December and work back because December is the most current month. However, most people do go through it fairly linearly. But anyway, this is just to give you an idea of what we're talking about when we talk about the event history calendar.

So the scope of the redesigned SIPP is very similar to what it used to be with SIPP classic. It's definitely broader than core, but it's - and it includes a lot of the topical module content. And here is just the table that kind of summarizes those differences. Again, now we have a Windows-based instrument and that of a DOS-based instrument. We still conduct the interview via personal visit with a little bit of telephone follow-up. That hasn't changed. We now

interview annually instead of three times a year. And therefore, we're asking about the previous year instead of the previous four months.

Panel length is about the same. We're anticipating roughly a four-year panel length. Sample size is a little bit larger. In the 2014 panel, we interviewed about 53,000 household. The survey universe is still the civilian, non-institutional population of the United States. And I'll talk a little bit more about what that means in a minute. There's still very comprehensive content, but the file structure is now much more simplified. And for those of you who have used SIPP in the past, that's a huge improvement. You are probably well aware of how frustrating SIPP used to be to use because you had to keep joining a bunch of files together. Now it's a much easier process.

We won't talk about it much in these webinars, but you should be aware that in between Waves 1 and 2 of the 2014 panel, we ran a supplement on behalf of the Social Security Administration. And that supplements collected additional data about marital history, pension receipt and disability. So if you happen to be interested in those three topics, I would certainly suggest that you also look into data from the SSA supplement. And that data, along with all of the SIPP data, is available for download at the census Web site or the SIPP Web site which is [census.gov/sipp](https://www.census.gov/sipp).

Okay, so I'll talk a little bit more now about the way that the survey is designed in a way that we designed the samples. So I mentioned before that the survey universe is the civilian, non-institutionalized population of the United States. So civilian means we don't include people who are in active duty. And noninstitutional means that if you're in prison or other sort of institutional group quarters like that, you're not included. And that's mostly because it's hard to get access to interview those people. It's also hard to

follow them and because SIPP is longitudinal, we are generally trying to follow people.

You should also be aware of the way we treat infants. And infants are defined as people who are younger than one year old at the time of the interview. They are listed on the roster, and they are available to be selected on other household members question. So if, for example, we're asking mom about her health insurance and it says, there's a question that says who else in the house is covered by this health insurance policy, she can select the infant on that. But we won't ask the infants about the infants health insurance directly.

The design of the sample, our sample is based on the 2010 Decennial Census frame. The 2014 SIPP was actually a first Census Bureau survey to use this updated frame. And that is drawn from the Census Bureau's Master Address File or the MAF, and along with other surveys when we pull addresses out then they are out of consideration for five full years. So in other words, if you're selected to be in SIPP, you won't be selected to be in the Current Population Survey or in similar household surveys for at least five years. And that's just to avoid having respondent get overburdened with multiple surveys at the same time. That is rare, but it can happen. And so we do this to address on duplication here.

So the SIPP sample is constructed as a two-stage, stratified sample. The first stage being that we select Primary Sampling Units. And then the second is that we select addresses within the Primary Sampling Units or PSUs. So what is the PSU? Well a PSU is the county if the county has at least 7500 people. If the county has fewer than 7500 people, then it's a combination of contiguous counties. And we have two kinds of primary sampling unit; self-representing and non-self-representing. And I'll explain what that means in a second.

But here's the visual example of what a primary sampling unit is. You can see that the - this Wyoming. These are not the actually primary sampling units, but this is just an example. You can see that each of these counties that is above 7500 people, is its own primary sampling unit, so Park County here with about 8,000, Teton County with about 15,000 and so on. And then ones that are below counties where there are fewer than 7500 people get combined. And so you can see that this primary sampling unit is a combination of three counties. This is a combination of two counties and so on.

And the color coding means that the purple are self-representing and that the white are non-self-representing. And what that means is that self-representing PSUs are guaranteed to be included in the SIPP sample. Non-self-representing PSUs may or may not be included. Basically we stratify them and then they are selected with the probability according to their size. And what this means in reality is that the 2014 SIPP contains a big chunk of 344 self-representing and 476 non-self-representing. And that means we're in about 1/3 of US counties and because the way the population is distributed, that third of the county covers a majority of the US population.

So continuing with our visual example again, you can see that here are some demonstration PSUs in Wyoming. Again, the purple ones because they're self-representing primary sampling units are guaranteed to be included. And then the ones that were in white were the non-self-representing, meaning that they may or may not be selected. And so in this example, you can see that we've selected cases in the green, primary sampling units and did not select cases in the white ones.

So once the primary sampling units are selected then we select addresses within each of those areas. And we stratified the addresses by income and poverty status and the idea with doing that is because we want to oversample



in low-income areas so that we can increase the number of poverty cases in the sample. Because much of what it measures is related to income and poverty; therefore, if we have more cases that are likely to be in poverty in the sample, we're more likely to get good data for those cases. And so all of that provides our final household sample which was, again, about 53,000 housing units in the Wave 1 in the 2014 panel.

I mentioned that we oversample in poverty areas and so how we do that is using American Community Survey or ACS data, we rank the housing units in each primary sampling unit into high income and low-income areas. And then select addresses within those areas. And basically that gives us low-income housing units at a higher rate. And so we got about a 28 percent increase in cases at or near the poverty level in the 2014 SIPP.

You should know that SIPP estimates are designed to be nationally representative and state representative. PSU borders do not cross state. So all of the sampling is done within the state. And in some states where we have a relatively small sample, we added sample to increase the reliability of estimates from those states. So ultimately, it is representative for the 2014 panel is representative for all states. That's not true of earlier SIPP panels. So if you're looking to do state-level research, be sure and check the documentation for that. But in 2014, you'll be fine.

So what does the sample of SIPP look like? Well it generally follows the distribution of the US population. And because the population in United States has more people in metropolitan area, we also sample more in metropolitan areas. In rural areas, we tried to get about 25 cases per non-self-representing primary sampling unit.

And to compensate for the way that the sample is designed, therefore we have weight that we include in the data set. And basically weights are the number of units in the population that the responding unit represent. In other words, if we sample one out of 2500 households, then the weights for each household would be 2500 to indicate that that household really represents 2500 households in the final population.

And again, we try to make it representative of the survey universe which is, keep in mind, the civilian, non-institutional population of the United States. So it won't quite match if you look at, for example, American Community Survey data which covers everything. Our estimates won't match that because our survey universe is slightly different, but they'll be very close.

So we have to use weight, or we could just using weight because otherwise you're going to have estimates that are somewhat biased. And that due to several factors including the sample design which I just described, the oversampling of low-income cases, as well as differential nonresponse.

And differential nonresponse means that different group members respond differently. And that can be geographically. So for example, people in the Midwest are more likely to respond than people in New York City. It's also age, older people are more likely to respond than younger people and so on. And so the weighting design compensates for all of that.

And we do provide weight at the personal level. And from that you can create weight at the family or household level depending on what your research question of interest is. You can also get weight at multiple timeframe. We have monthly weight. The December weight you can also use as a calendar year weight. And then finally in Waves 2+, we have a panel weight. So again, you want to choose your weight based on what your research question is and

how you're designing your study. But really for whatever you're doing there should be a weight that fits your needs.

The way we construct the weights, here's sort of a graphical representation. And I'll walk you through that in a minute. But essentially you start with a base weight. It gets adjusted by various things and ultimately, you know, each of these areas represents an adjustment and then ultimately you get to the final weight.

And so we start with the base weight. And the base weight is what I mentioned a second ago which is the inverse of the probability of being selected. So if, for example, a county has 50,000 household and we sample 20 of them, therefore each one is going to have a weight of 2500 representing the fraction of the population that got sampled. Then we adjust the base weight. So first there's duplication control factor. Then there's a nonresponse adjustment.

And then there's a second stage adjustment. The duplication control factor means that occasionally when we get to a sample address, we find out that instead of the one housing unit there, there are multiple. So if for example, we thought it was a house and it turned out to be an apartment, we would need to subsample to make sure to get only one interview for that sampled address. But because you do that, that will then change the households probability of being selected. And so the duplication control factor adjust for that. Again, that's relatively rare, but it does happen and so therefore, we have to adjust for it.

Next we have a nonresponse adjustment and that is just for what I said a minute ago which is that not all household that are eligible are going to be interviewed and they're not going to be interviewed, then non-interviews don't

happen at random. They vary from group to group. So in this example continuing what we had said before, we sampled 20 household in that county. If only 16 of them responded, then they got adjusted by a factor of  $20/16$ . So instead of the base weight of 2500, then it got adjust to 3125. And this example, sort of, implies that the missing this is at random, but as we discussed, it's not. So we do some additional nonresponse adjustment to account for that.

The final adjustment is the second stage adjustment. And the second stage adjustment means that we want to make sure that our estimates agree with independent monthly estimates of the population. And so for Census Bureau, the things that we look at are, among others, state, age, race, sex, Hispanic origin, family relationship, and so on. And this adjustment is actually done at the person level rather than the household level because, of course, the household can have a mix, people with a mix of these characteristics. So it's not really possible to do that the household level.

Anyway, after all of that adjustment, you get your final person weight and that's what's available to you on the file. And as I mentioned from the person weight, you can then derive a household or a family weight depending on the needs of your research.

So from those things then we have a sample and we're ready to go out into the field. And once we do that, here's kind of an overview of what we ask respondents and what we collect. Before we get into that, I want to go over a few definitions just to, sort of, make sure that everyone is on the same page and that when I refer to these concepts, they'll make sense to you.

So first we have a respondent. A respondent is simply somebody in the household for whom SIPP collects information. And for SIPP that's basically

everyone in the household. Everyone who's actually living there. If you're just there visiting, you won't get counted as a respondent. But if you live in the household, you'll be a respondent.

Specifically there are also household respondents and household respondents are the first eligible adult household member who is interviewed. And that's often the same but can be different from the reference person. And the reference person is the owner or renter of the housing unit. And if there are multiple people on the lease or multiple people on the mortgage, it's the first of those people that we talk to.

The reference person is an important concept because a lot of the subsequent questions are asked only of the reference person or they're asked about other people in the household in relation to the reference person. So just keep in mind that's the owner or renter of the housing unit.

Type 2 person, this is something that is new to the 2014 SIPP and a Type 2 person is someone who lives with a respondent during the reference year but is not living there at the time of the interview. So if for example you had a roommate over the summer and the roommate's no longer there, the roommate would be a Type 2 person. We're not going to interview the roommate. We're going to collect very minimal information about this person. The important distinction is that a Type 2 person does not have a person record. They are on the information of the person we did interview who reported the Type 2 person. They do not have their own records in the dataset.

And also keep in mind that because we collect a year's worth of resident history for each respondent, a Type 2 person could have lived with the respondent at any address where the respondent lives, not just the address at which they got interviewed. So again, using the example of the person having

a roommate earlier in the year, let's say, the respondent had actually lived at the, in a dorm or at the house with a roommate, we would collect that information and then the respondent moved to the interview address and there are no Type 2 people there. But there were at the previous address.

So for a lot of research, you may or may not be interested in what the Type 2 person is doing, but for a lot of research that involves family dynamics or income or anything like that, you'll want to include the measures with the Type 2 people so that you're getting a fuller sense of the household's economic and social circumstances.

SIPP defines adults as anyone 15 years or older and so keep in mind that a 15-year-old child who's living in the house is going to get asked questions about fertility, is going to get asked questions about employment or marital history. And for a lot of 15-year-olds, they're not working. They're not married. So it's not relevant to them. But sometimes they are. So even though people tend to think of adults as 18 or older, keep in mind that for SIPP they're defined as 15 years or older.

SIPP like the Census Bureau in general defined family as a set of people related by blood, marriage or adoption. And the household is defined as a set of people who are living together. And then the next definition is self versus proxy and we'll, you may hear us refer to these kind of interviews throughout the webinars. Basically as self-response is that the respondent is providing information for him or herself. And a proxy response is that another adult in the household is providing information for the respondent. So if we go to somebody's door and come you know, it's a married couple, husband and wife, only one of them is home. That's spouse can report for the spouse who's not home.

And for all children, i.e. everyone under 15, those interviews are all done as proxy interviews. So adults are always providing child interview information. We never interviewed children directly.

And finally we have the coverage unit. And this is also known as the program unit. We will use those terms interchangeably. And that's a set of people who are covered by a health insurance policy or an assistance program. So for example, you know, your food stamps may cover you and your child, or you and your spouse and your child. Whatever group of people is covered by those food stamps is referred to as the coverage unit.

And similarly with the health insurance policy, whoever is covered by that policy is part of the coverage unit. So that could include one person. It could include some of the people in the household, all of the people in the household. It may even include people in the household as well as, you know, health insurance policy might cover, for example a kid who is away at college. That kid would be part of the coverage unit even though they won't be interviewed as part of SIPP.

So now that we have those definitions out of the way, here is a high-level overview of what the survey asks of respondents. And we'll go through - the subsequent slides have this in more detail. We start by collecting a roster with everyone in the household and we collect their name, their sex as well as their birthdate. We collect demographic information about each person and that's the standard Hispanic origin, race, citizenship, language, English speaking and so on, their marital status and additional demographic information.

Then we get into the event history calendar like I showed you before. And that covers a variety of topics including residency, marital history, educational enrollment, jobs, as well as, time spent not working, program receipt of a

variety of assistance program and health insurance. After we complete the event history calendar, we have some follow-up questions for health insurance, some additional questions about kind of assistance such as dependent care, non-job income, other income that you might have received from programs. We had some detailed questions about asset ownership and household expenses. And then we ask them questions about healthcare utilization, medical expenditures, disability, fertility history, and then some childcare and adult and child well-being questions.

And then finally we close out by asking about contact information for each person because we're going to interview them again the following year as well as whether they intend to move. And of course not everyone knows that they're going to have moved by the time the next year rolls around but if we know that they are planning to move, that will help us locate them when we go to find them a year later.

Okay, so here's some more detailed slides that show the questions for each of the subject matter areas. So we start with demographics. And again in demographics we collect a roster of all the people living in the household. We then collect a roster of the Type 2 people which you'll remember are people that lived with you during the year but are not currently in the household. Some standard demographic information about each group, about each person. We collect detailed information about household relationships including options for same-sex partners and spouses. And things like marital status, and then a brief marital history of people. So the year that they entered their current marriage, the numbers of times they've been married overall, whether been widowed or divorced and so on.

Next we ask for all children in the household. We asked a gender-neutral parent identification so that again, you can identify same-sex parents or



opposite-sex parents. Basically it points to Parent 1 and Parent 2. The type of parent so whether it's a biological child, a foster child, an adopted child, a stepchild. For each person we collect some nativity and citizenship information, some educational attainment information, later in the event history calendar we'll collect educational enrollment. That's current educational enrollment.

But here in demographics, we collect educational attainment as well as things like educational certificates, professional certifications and licenses and so on. And then we collect some information about people armed forces status, whether they're veterans, whether they're on active duty, and so on.

After demographics, then we get into the event history calendar. And this is the same graphic I showed you a little while ago. Again, the months across here, so we'll try to fill in each of these months for each of these topics. And of course, not every topic will have coverage in every month. For example, you might have only been on food stamps for a few months. You might have had health insurance for part of the year but for a few of them, marital history will collect the whole year's worth and residences will collect the whole year's worth because even if you didn't have a fixed address, you were living somewhere during the year. So we will collect that information.

So again, for residences, we'll collect up to five residences for each person. For each of those residences, we'll collect a tenure status meaning did you own or rent the house. Then we ask whether you received any public housing assistance whether you got a subsidy, whether you got a voucher. We asked why you moved to the address. And then for the address that began in January, we'll ask when you moved into that address and whether, what your tenure status of the prior residence was. That will allow you to look at tenureship. So did somebody go from renting a house to owning a house? And

things like that. So for mobility researchers, this is - there's a tremendous amount of information here.

The next topic in the event history calendar is marital history. You can report up to three spells during the year. And what that will give you is monthly marital status with a pointer to the spouse if the spouse is in the household, cohabitation status, if you're not married but you had a partner in the household, and then the registered domestic partner if you're cohabiting, you can say that you had - you were in a formal registered domestic partnership.

Next is educational enrollment and you can report up to three spells there. And for each of shoe spells, we will collect the grade that you were attending, the type of school you're in, so public private, and so on, whether you enrolled part time or full time, whether you had to repeat a grade and for kids who are seven and under, whether they were in head start.

After educational attainment, we go into the labor force section. And this is a very thorough section as you can see. And we collect information about up to seven discreet jobs and you can report up to two spells working in each of those seven jobs. And if you happen to have more than seven jobs, you can fill in a timeline of additional work. So for the jobs beyond seven, we don't collect information that's as detailed, but we at least collect what - when you were working in those jobs.

For each of the seven jobs, we do collect the kind of pay that you got as well as the amount, your earnings if it was a job or your profits or losses if it was a business, how many hours you worked per week, whether those earnings or hours changed and we can collect up to three changes, the industry occupation and class of worker, the kind of information, some information about your business or if you have an employer, and that's things like the address and the

size of the business and so on, whether the person was in the union, whether the business was incorporated and then whether you had any time away without pay if you were looking at gaps in people's work history.

For people who aren't working, we collect spells of time out of the labor force or time not working. And those are separate because you could be out of the labor force in that you were going to school or you were retired. Or you could be in the labor force and looking for work. And so we ask some question to differentiate which of those two groups people fall into. We ask why they weren't working, whether they were available for work and if not, what their reasons were.

For each of those jobs we also collect information about people's commuting and work schedule. So for commuting we ask how people got to work and you can report multiple modes. So if some days you take the bus, other days you drive, you can report that. We ask how long it took you to get to work as well as how far it is and about some cost. And then as far as work schedule goes, we ask each days of the week that you worked, when you started and stopped stay, whether you worked from home, and what kind of schedule. And the kind of schedule is daytime, nighttime, shift work and so on.

After work finished the job, then we move into the program section of the EHC. And the EHC collect data on multiple programs as you can see here. So that includes SSI or supplemental security income TANF, temporary assistance for needy families, supplemental nutrition assistance programs, as SNAP which is also known as food stamps, general assistance and WIC. And for each of those programs, we collect up to three spells. We asked who in the household is covered as well as who owns the coverage. And those can be different depending on the way it's set up. So for example, perhaps you get food stamps, but you're not covered but your children are. So you might own

the coverage, but you're not covered. And so these questions help differentiate that. We ask why you started and stopped receiving the program as well as the amount that you received.

After programs, we have some health insurance related questions. And we ask these questions separately for private health insurance, Medicare, Medicaid, military health insurance and any other coverage you might have. And similarly to programs, we ask for each health insurance policy, who in the household was covered, who owns the coverage, whether anyone outside the household was covered, what type of coverage it was, so whether it was an HMO, PPO or so on, how much it costs you and whether it was a public or private plan and what kind of deductible that you have. And for Wave 2+, we've added some questions about the Affordable Care Act. So he asked whether people obtained their health insurance coverage via a state exchange and a couple of other questions. But that started, keep in mind that started in 2015, so it's not on the Wave 1 file.

So after finished with health insurance, that wraps up the event history calendar. After you exit the event history calendar, you get some health insurance follow up questions. These help to reconcile any time that people had without coverage, if they were employed but they didn't have private coverage will ask what the reasons were for that as well as any other reasons they were covered either by public insurance or by no insurance at all.

So after the health insurance follow up question, then we have some additional questions related to annual programs and these are various types such as Social Security income, VA benefit, workers comp, unemployment comp, energy assistance, whether you were - your children were enrolled in free and reduced-price lunches or breakfast or lunch is at school, whether you receive any other of these kind of income here, whether you got child support,

alimony and so on. And then finally some follow up questions about other kinds of non-cash assistance, cash and non-cash assistance such as training, food, clothing, housing and so on.

After the program section, the next section of the instrument involves assets. And the assets ask both similarly and jointly we collect - jointly meeting, you know, you own it with your spouse or with somebody else, for each of those will ask balances and values for all of these kind of assets.

So everything from checking and savings accounts to mutual funds to stocks to retirement income, or and then some less common kinds of acts that put out as royalties, you know, mortgages that you have as an investment, so very detailed acid questions and then we asked some questions about other kind of nonfinancial assets such as real estate, vehicles that you own. We collect information up to three vehicles in the household and we ask about the newest vehicles because those are the most likely the more valuable vehicles that you would own.

And then if you own a business, will ask about the value of the business and any debt the business and has as well as retirement balances and any unsecured liability which are things like credit card bills or medical bills. We also ask more generally about medical expenditures and those include things like your out-of-pocket expenses as well as how you utilize healthcare.

So did you go to the doctor, were you hospitalized, how many times, how many sick days you had to take in the previous year, what kind insurance premiums you faced. And if you are uninsured whether you are able to go to the doctor at all.

We do ask some disability questions and these questions are consistent with disability items that are on other Census Bureau surveys so you may be familiar with them even if you're not familiar with SIPP. So everybody gets asked whether they have some sensory disabilities which are problems with seeing and hearing. Adult get ask these questions here.

So problems with concentration with memory and so on. And in children get a similar but slightly different set of questions. Again, concentrating and memory, but then also do they have issues with playing with similarly aged children. Do they have issues during their schoolwork? And in very young children get a questions about development conditions or delays.

And I mentioned at the beginning of the presentation, we did a supplement on behalf of the Social Security Administration. And that supplement does include detailed questions about adult, child and work disability. So if you're a disability scholar that SSA supplement is probably a very good resource for you to use in conjunction with SIPP data.

After disability, we move into asking about fertility. So we collect to complete fertility history of all adults in the household. And so we get the roster in ages of children who you've given birth to or your father. And what that enables you to do is construct a measure of multiple partner fertility. And that's something that is not available in most other surveys. This is a real strength that SIPP is that you have this indicator of multiple partner fertility. You also get an indication of whether the respondent is a grandparent or not. And then for each respondent's biological parent, we ask them questions about nativity, their nativity and mortality including the country of birth and whether they're still alive.

Then we move into asking for each child in the household about their childcare. And so we ask what kind of arrangements and then for each kind of arrangement, and then which children, child or children use those arrangements. How much you had to pay out-of-pocket, whether you received any assistance with those costs, and whether you had to lose any time from work because of childcare.

And then the final set of questions are about well-being. And these are different whether the respondent is a child or is an adult. But the children get the battery of questions about whether they had dinner with her parents and whether they were engaged in school and so on. And adults get questions about the environment. So is the neighborhood on say? Was it noisy? Does the housing unit have problems with path? Does the plumbing have problems? Or does that household have any problem with their ability to pay the mortgage or the utilities?

That leads into the food security questions which are again a sort of standard battery of questions that you may have seen on other surveys. And they ask whether the household was able to buy enough food, whether they were able to eat balanced meals, or maybe they had to cut back on their meals. And if so, did they - how frequently and then did adults defer food to the children?

So those are the, all the substantive kind of questions that we ask. Of course, at that point, you know, we've been in the respondent's house for a while and they're ready for us to leave. So we have a few follow-up questions here just sort of asking again I mentioned earlier, you know, about moving intentions. And we got some contact information for other people that might be able to tell us where the respondent is if we can't find the respondent.

So then we close out the interview. And so at that this point we've collected data for the household. And once we get the data back to headquarters, then we have to process it before we can get it out to you guys so that you can analysis it. And in order to do that, we have to do a number of things and I'll sort of, walk you through what those are.

So with the 2014 panel besides the fully redesigned instrument, we also fully redesigned our data processing system. And that took a tremendous amount of time and effort on behalf of our subject matter and our programmers. Not least of which was that they had to convert the editing programs from four tran to SAS which sounds straightforward but was a tremendous effort. But the basics of our strategy and processes, our processing strategy do remain the same. And that strategy is mostly what we want to do is item level imputation.

So that occurs when a respondent might be in the universe for a given question but has not provided an answer. And that could either be because they don't know. It could be because they refuse to answer. Or it could be there was a technical glitch and they just didn't get the question. That's relatively rare. It's much more likely that they don't know especially with a proxy interview, but it could also be that they refuse because they don't want to tell us.

But we need to fill in that missing information. And so we do that via one of three ways. We can either do logical imputation, hot deck imputation or cold deck imputation. And these are listed in order of preference. So we prefer to do a logical imputation. And that simply means that we use the respondent's answers to other questions, related questions, to derive the answer to the missing value.



So for example, this is not a real example. It's an oversimplification, but here's an example. Let's say the respondent has three children and for the middle child, for whatever reason, they didn't tell us what - where the child was born. But for the other two children, they said that their kids were born in Texas. So logically the missing value may get set to Texas. That's not foolproof, but it's pretty good given people tend not to move and so it's a pretty good bet that the missing value is Texas.

We also do use this method to edit, provide the answers for logical consistency. If for example, the respondent says they were born in the United States, but they're not a US citizen, that's not possible. That's logically inconsistent. And so we would most likely set them to have been born in the United States and to be a citizen.

So that's logical imputation. If for some reason we can't do logical imputation, then we'll move to hot deck imputation. And hot deck imputation uses, and answer provided by a very similar respondent to a missing value. So if, for example, the respondent doesn't tell us whether they owned or rented their house, but similar respondents do tend to own their residences, we may set the missing value to owned. And similar respondents are matched on a variety of characteristics, usually demographic characteristics such as age, race, educational attainment and so on.

And this has the advantage of that the missing data will reproduce the distribution of the reported data. So we're not going to skew estimates with this imputation method.

And finally if the other two don't work, then we will use cold deck imputation. And cold deck imputation just uses a default value to be inserted into the missing value. We don't ever really want to do this, but if we do have it there

as a last resort. And the cold deck value usually is just the most common answer for the question. So again if it's owned or rented, most people tend to own their houses, so we would use that as the default answer.

So these methods work when you've got a single missing item, but what happens when there's an entire person who's missing. And in that case, we have to use whole person imputation. And in previous SIPP panels we would just take a similar person, again matched on various demographic characteristics and just take all of that person's information and substitute it for the missing person. If you've used it before, you may have heard that referred to as Type Z imputation. That's the same thing.

However, that's a fairly crude method of imputation. And so for 2014, we've refined that, and we've replaced the whole person imputation with model-based imputation. And basically this uses statistical models to provide some missing data for the respondents. And the models are constructed using fairly sophisticated techniques that the - are method areas have developed. And it's the combination of survey data and administrative data and essentially what that allows us to do is to bring in data from outside the pool of survey respondents. So instead of just to model the data with the relatively small number of SIPP respondents, we can model it from a much larger population and therefore the model is more precise.

However that leads to a problem. How can you create models when you're trying to get data out for a survey that has 11,000 variables? There aren't that many variables on the public use file, but for the imputation to work properly, that's the number of models we'd have to come up with. Well if we were to do that, you know, we'd be creating models from here until kingdom come.

So we decided we're not going to create 11,000 models. What we're going to do is model what we call topic flags. And essentially, topics like our indicator variables for each of the major topic areas that we cover. So the model will say yes, this person had a job or know this person didn't have a job. And then we can use the regular logical and hot deck imputation on the more specific questions.

A little bit more about topics like, so as you seen, this instrument is divided into different subject areas. And at the beginning I must of the subject areas there are one or two. A question that determine whether the respondent should be ask a more detailed follow-up question for that topic.

So for example, at the beginning of the job section, the first question is do you currently have a job or business. And if the respondent said yes, great, then they go into the detailed job questions. If the respondent says no, I'm not currently working, then the next screen of questions that okay, did you have a job or business at any time during 2013. And again, if the respondent says yes, they go into the detail follow-up questions. If they say no, then they moved to the next section.

So what to be modeled been those is it tries to just fill in this. And so what we see when we look at the topics set's a one, if the respondent had a job. The set it's a zero if the respondent didn't have a job or the they're missing if the respondent skipped the topic entirely. And it's these missing values that the model will try to fill in.

And so what that does is it lets us stop whole person imputation. It also preserves the correlation across topics because you can estimate a joint distribution. And what that means is if you're missing say both job and health insurance information for people, those two things are often connected. And

so by having the model take both of those things into account, the estimates will be more precise. And you can estimate the likelihood of health insurance contingent on whether the person had a job or not as opposed to just imputing them separately. And if you do that then it doesn't look like there's any relationship between those two even though we know that there is.

And so that the advantage of that, then we're only imputing a relatively small number of these topic flags and then they can be used in the downstream edits. So the topic flag then whatever value is imputed there if it sets a yes, then the item level imputations can then set the values, the missing values for the detailed follow-up questions and if it sets a no, then we can just go to the next subject.

So basically here's a list of what topic flags get imputed . So within the event history calendar it's educational attainment, whether the persons worked or not, whether the person received each of these assistance programs or not. And then whether the respondent had each of these types of health insurance or not.

And out outside of EHC we, the model will impute things like whether respondent had any children, yes or no, whether they were disabled, yes or no, whether they were receiving retirement income, yes or no and so on. And so essentially it's a series of yes/no questions that the model predicts.

And here's just an example of what this looks like. So of the people who were in universe to be asked about whether they work for pay, about 95% of people responded to this. And of those 95%, about 58% said yes, they worked and about 42% said no, they didn't. And you can see that outlines that very well with administrative data that we get from the IRS. Even though it's lagged here, these IRS data are for 2012 and we were asking respondents

about 2013. You can see that it's very consistent. The percentages are almost dead on.

For the 5% of people that did not help us whether they worked or not, in other words they should have answered the question but did not, you can see that the model imputes about 62% of them to have worked for pay and about 38% who did not. But if you're comparing it to the reported percentages, you say well the model is overestimating people that work. Well, the reason it does that is because of differential nonresponse.

So in other words, people who do work are less likely to tell us about their jobs or their lack of jobs has been people who are not working. So in other words, if we were just using logical imputation we would have reproduced this distribution, but the model is a more precise distribution of the people that should have responded because it lines up better with the administrative data.

So we've done a number of these evaluation and the model definitely produces more precise results. So we think this is a big methodological improvement within the 2014 set. And we're happy that we've managed to expand the processing system to include this.

So beside the topic flags, another kind of like that we create in the processing system are status flags. And if you use it before, you may have heard these referred to as allocation flags. They are a similar idea, but now they contain a lot more information. And so for each of the variables in the data sets, you have a matching status like. And it's denoted by an a in front of the variable name. So for example the E tenure variable which met - which indicates whether the person owned or rented a household, there's a corresponding

status flag called A tenure. And that's the same for every variable in the data sets. Every E variable will have a corresponding A variable.

And there's a lot of information in the status flag. You can see maybe the person just wasn't in the universe for a given question, maybe they were in universe and we took the reported data, maybe we had to impute the data, or maybe especially for recodes, it's a combination of these. And so these seven to nine are used for our recodes and I'll talk a little bit more about recodes a little later. Anyway so you can get a tremendous amount of information using the status like.

And then the final kind of flag that we create in processing is what we call a continuation flag. And you need a continuation flag because in the event history calendar we collect data throughout - through the month of interviews. So if we interview you in March and we're asking about the previous calendar year, we're collecting data from January of the previous year through March of the current year, so 14 or 15 months. But were only releasing the data to the end of the referenced year.

So as a result of that, you just looking at the data arts going to know whether the person's spell ended in December or is ongoing. And that can be important if you're looking like jobs, you need to know whether the job really ended in December or if it's ongoing. And so the, for each of the December months in the history calendar content we do provide these continuation flag. And what the show is that the spell either it ended after December, whether it ended after December but before the interview or whether it was ongoing as of the time of the interview.

So you can use these to see especially if you're doing like a survival analysis or an event history analysis, you can use these continuation flags to see

whether the person encountered the event. So at this point all of the processing is done and now we release the data.

And so I want to go a few things that you should know as you're working with the data or as you're getting ready to work with the data. So keep in mind, it's important to keep in mind what SIPP is good for and what it's less good for. So SIPP is very good for generating monthly, national-level estimates as well as the 2014 panel state-level estimates. It's good for annual estimates. It's good for longitudinal analysis over the length of the panel keeping in mind that the panel is a four-year panel. And because we oversample in high poverty areas, SIPP is very good for analyzing income and other economic circumstances of four people.

It's not as good for long term longitudinal analyses. Again, the panel is four years long, so it's not going to cover anything before or after the four-year period. And it's not as good for analyzing data at the top of the income distribution because we don't have nearly as many wealthy or high earning households in the sample as we do poor households. So just keep that in mind as you're thinking about your research question and thinking about how to use the data.

Just a few tips, our datasets are provided in SAS and ASCII format. They're available on the SIPP site and on the FTP site. Starting with the 2014 panel, we've started making the files also available in Stata format. We know that a lot of people are now using Stata and so we're trying to make the files available in user-friendly formats like that. If you have a SAS file and you want to convert it to Stata and you don't have a tool like transfer, this Web site that UNC maintains is a good tool for moving datasets from SAS to Stata and they've got a parallel command that goes in the opposite direction.

Keep in mind as you're using the dataset, SIPP has a lot of variables and a lot of observations. Because it's a person month file, keep in mind that each person has 12 observations and there are about 2500 hundred variables on this public use file plus 2500 correspondent status flags. So the file is large and depending on your machine, that can lead to significant memory limitations. So it's definitely worth working on a well-equipped machine if you have access to it. And when you load in the dataset for speed and performance reasons, we definitely recommend that you keep only the observation and variables that you need so that you're not executing your commands over a much larger dataset and slowing yourself down.

We also recommend that you keep your complete SIPP files in their original state and don't change or save on them. Save any subset files you need as separate files rather than updating your master files. We recommend that you create a single SAS program or Stata do file to construct the dataset and then, of course, you can always just reconstruct as you need to.

Just a few best practices, again, you want a permanent set of programmer do files for your analyses. First you can have one that constructs dataset. Then you can have one that constructs the variables. And then finally you've got some analyses that you can run. And if you do this, of course, it'll make it a lot easier for you to add variables to your dataset or take them out. It's much easier to find mistakes because it's debugging is simplified a lot. And then of course you need if to rerun your analyses, you can just do that pretty quickly.

I talked a little bit before about the variable scheme and so it's very useful to keep in mind if you're looking at the data, how the variables are named. So E variables are simply edited versions of reported variables. And so those are things that we ask, and respondents answer directly. So tenure, did you own or rent the house, male or female, and so on. The A variables are, again, the



status flags that I mentioned and again, keep in mind there is a status flag for each of the variables.

You may also see variables that start with T, and those variables are top or bottom coded if they're continuous variables or they're collapsed if they're categorical variables. And this is usually done for confidentiality or disclosure avoidance reasons. So age is top-coded because there aren't many people above say age 90 in the SIPP panel, so everyone above age 90 is top-coded. I think we top code at age 88 or something like that.

So that's a continuous variable that gets top-coded. T LIV QTR is living quarters type, and this is a categorical variable that is collapsed because most people live in type value of one which is house, apartment or flat. And so for the other types which are everything from, you know, a trailer or it could be an unoccupied tent site where you were camping for a while or whatever, those get collapsed down because there aren't that many people who fall into those categories.

The final kind of variable are recodes, and those are denoted within R. And recodes are variables that are not directly asked to respondents, but they're created during processing from other variables. So this is a language isolation recode which measures how well respondents speak English. The R mover recode is created from the residents history information and indicates the kind of move that a respondent undertook. So whether they moved within a city, across a county line, across a state line, and so on. So the recodes provide a lot of additional information that again we don't collect directly from respondents, but that we can derive.

Also keep in mind if you're, when you're using the data that you may need to identify unique respondent, and this is because it's a person month file there

are 12 observations per person per wave. And so you'll need to have a person identifier and you can do that here and then you can very quickly if you are only looking at the December for example very quickly drop the other 11 observations for that person this will allow you to identify those people. And then in subsequent webinars, we'll discuss how you can identify families because they're not noted directly in the file.

I mentioned you want to keep only the observation that you need. If you're looking at individuals, you want to keep all the respondent observations. You may only need one per person or you may need all 12 depending on the kind of work that you're doing. If you're looking at households, you probably only want to keep one observation per household and again, you probably want to use the head of the household or the household respondent and you can figure that out by using the relationship question that we'll discuss more in the demographics webinar.

Again and we'll discuss creating families, but here's sort of a very high level. We recommend that you keep one observation per family. The advantage is that for household and family level variables, we record those on each of the sample members observations, so it makes life a lot easier. So every member of the family will have the same value for these variables or every member of the household.

Because SIPP is longitudinal, you should also think about how you're going to follow people across waves. So in the survey, basically everybody who's there in Wave 1 is considered an original sample person or OSP. And in subsequent waves we will attempt to interview every OSP as well as any new people who have moved in. So if, for example, in Wave 1 there's a married couple. In Wave 2, they've gotten divorced. One of the spouses has moved out. And then has a new partner living with them, then we would interview

both the original spouse, the spouse who moved out as well as the new partner who moved in with that spouse. So then all three of those people would be in the survey universe.

And for original people who leave the household we'll call them as long as they remain in the survey universe, which is again the civilian, noninstitutionalized population and in the United States. And that is denoted in this variable called T HHLID status or T-household status. And so a value of three indicates people who were interviewed in the previous wave who have since moved out of the survey universe. And that can happen in a variety of ways. They may have moved into the institutional group quarters meaning that they might be in jail. They might be in a nursing home. They might also have moved into military barracks or they're on active duty. They may have moved out of the United States or they may have died.

And for these people we'll collect a limited set of information about them and that will allow - that allows us to impute records for them for the month that they're in the household. But of course months after they've left the household they're set to inactive.

Similarly, we have household status equals four people and those are people who were interviewed in the household in the previous wave. They have moved out of the household, but we've been unable to find them. So they may be in the survey universe, but we just haven't been able to locate them. So these people get set to inactive, but they are available for selection on rosters of active people. And as I said before, those can be people who are like again if you're saying who does your health insurance policy coverage. It covers me and you can also select people who are no longer active as having been covered by you.

So if you're doing research across waves and you want to link multiple waves to the SIPP, also think about which file structure that you need. Because you could combine the files in a wide structure, or you could combine them in a long structure. And ultimately, that's going to depend on what your analysis is and how comfortable you are working with data. This is somewhat of a consideration when you're working within a single wave, but it's critical when you're working across waves.

And what do I mean by wide and long files? Well a wide file is a file that has all of the waves variables put on a single record. So in other words, this is a January record for Person 101. And you can see that their Wave 1 age is 35. Their Wave 2 age is 36. So in other words it's wide because it keeps extending out this way. You can keep adding multiple waves to this file.

Conversely you might want a long file where instead of going out this way, you create, you join the two waves of data and then instead of 12 person months, you now have 24 person months for a certain individual. So in other words, you would have 24 observations for each person.

So we have given you some sample code, how do to do this. If you look at the Census Academy site, if you've downloaded the PDF that has the workshop slides that I'm going over right now, at the back of that I some sample code both in SAS and Stata format about how to create either wide and long data. And I thank (Holly Fee) for creating that same code for us.

Okay so once you've decided all of that, you also need to keep in mind which weights you want to use as we discussed earlier. You don't have to use weights, but we strongly suggest that you do. You may actually want to run your analyses both ways, both with and without weights. And just see how much it makes your estimates differ. Sometimes they're very similar.

Sometimes they change dramatically. But certainly for any kind of work that you're going to publish or any research that you're doing say for a project, school project or a dissertation or whatever, we would definitely recommend that you use weights because that compensates for the way that the sampled is designed.

All right, so that is the end of some things to keep in mind with the data. Now just a couple of resources for you guys as you're participating in this webinars. As I mentioned, we have the presentation slides and sample code currently available. You can access it at the Census Academy site here. That's up there right now. Later on will be a recap of this webinar, but you can get the materials right now. If you want more detailed data, your best resources are these sites. There's the SIPP site which is [census.gov/sipp](https://www.census.gov/sipp). This is a great resource for all things SIPP. The Census FTP site is good if you just want to download datasets.

And the third site is not a sensitive sponsored site. The National Bureau of Economic Research maintains a very good SIPP page. And they have a lot of datasets. They have the same datasets that we have, but they've converted them to other formats including SPSS and Stata for panels prior to 2014. They also have a lot of good sample code and a lot of good frequencies and things like that. So like I said, although this is not a Census-sponsored resource, this is a place that we often send data users because it's a very good resource.

As I mentioned [census.gov](https://www.census.gov) is a great resource internally. We have a lot of documentation that we've made available up here in addition to the data. There's a very good user's guide. There's metadata and data dictionary as well as some, you know, very specific notes to think about or things to think about if you're using the data. We have a codebook. And then we've create crosswalks if you've used the 2008 panel and you're curious about what's

changed. These two crosswalks will help you compare 2014 and 2008 and you can get a sense of any updates you would need to make to your own code to run it on 2014.

So that concludes our webinar today. I will remind you that tomorrow Holly is doing demographics and residences and then later on in June we have another five or six webinars that cover the different subjects including jobs and assets and so on.

And certainly we encourage you to attend any of those that piqued your interest and that is it for today. Thank everybody for attending and listening to us. If you will have plenty of time for Q&A now. If you want to follow up with us later, you can send an email to [census.sipp@census.gov](mailto:census.sipp@census.gov). You can give us a call on the SIPP line or again, the SIPP Web site is there. So thank you very much. And we will now take questions.

Coordinator: Thank you. We will now begin the answer and question session. If you would like to ask a question, please press Star followed by 1 and record your name clearly. To withdraw your question, please press Star 2. One moment please while we wait for the first question.

Matthew Marlay: I saw that somebody in the chat asked for the URL for today's slides. They are available here on the Census Academy site.

Deborah Rivera: (Marcus) do we have any questions in the queue?

Coordinator: We show no questions in the queue.

Matthew Marlay: All right, well thank you guys very much. And again if you have questions later that you would like to follow up with us directly, feel free to contact us either via email or on the phone.

Coordinator: I'm sorry, two questions have just popped up int eh queue.

Matthew Marlay: All right.

Coordinator: One moment. Our question comes from (Chad). Your line is open.

(Chad): Hi, early on in the presentation, you had talked about the limitations of analysis below the state level. And it was just a quick touch on it. And wondering if you could elaborate or if that's something that's going to be in a subsequent session.

Matthew Marlay: It's not in a subsequent session. So sure, I can talk about that. So for SIPP, the sample size is not such that it can produce reliable estimates below the state level. Plus unless you've got access to the internal files, you're not going to have the geographic identifiers to do so. So we recommend that you use SIPP only for national and state-level analyses.

(Chad): Thank you.

Matthew Marlay: Our next question comes from (Danielle). Your line is open.

(Danielle): Thank you. Due to survey fatigue, are there certain portions of SIPP that tend to be skipped more often than others?

Matthew Marlay: So we have looked at that. And it is definitely the case that people break off, but it doesn't - what our field representatives were concerned about for

example was that because the nativity and citizenship stuff is at the very beginning of the survey, that was causing respondents to break off. And we weren't seeing that. And I have - Holly and Shelley can correct me if I'm wrong, but I don't recall having seen anything where there's like a general specific breakoff point where we can say, okay, by the end of, for example, assets people are sick of it. And are done. It seems like the breakoffs are spread pretty well, pretty evenly throughout the survey.

So I think, so my guess is that respondent fatigue is more in that, if it's a large household we might get interviews for the first two or three people and not, you know, People 4 to 6 as opposed to a certain person within the interview breaking off.

(Danielle): Thank you.

Matthew Marlay: You're welcome. Are there any other questions?

Coordinator: We show no further questions in the queue.

Matthew Marlay: All right, well thank you very much everybody. We appreciate your attendance. We appreciate, you know, your support of SIPP and your interest in it. And again, if you have any other follow up questions, feel free to contact us using the information shown here on the screen.

Deborah Rivera: Thank you, Matthew, for a wonderful presentation.

Matthew Marlay: Our pleasure.

Deborah Rivera: Appreciate it.



Coordinator: Thank you for participating in today's conference. Please disconnect at this time.

END